SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Computational life sciences
• RESEARCH PAPER •

November 2014 Vol.57 No.11: 1115–1120
doi: 10.1007/s11427-014-4735-x

Inferring microbial interaction networks based on consensus similarity network fusion

JIANG XingPeng[1,2] & HU XiaoHua[2*]

[1]*College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, USA;*
[2]*School of Computer, Central China Normal University, Wuhan 430079, China*

Received May 15, 2014; accepted July 21, 2014; published online October 17, 2014

With the rapid accumulation of high-throughput metagenomic sequencing data, it is possible to infer microbial species relations in a microbial community systematically. In recent years, some approaches have been proposed for identifying microbial interaction network. These methods often focus on one dataset without considering the advantage of data integration. In this study, we propose to use a similarity network fusion (SNF) method to infer microbial relations. The SNF efficiently integrates the similarities of species derived from different datasets by a cross-network diffusion process. We also introduce consensus *k*-nearest neighborhood (C*k*-NN) method instead of *k*-NN in the original SNF (we call the approach CSNF). The final network represents the augmented species relationships with aggregated evidence from various datasets, taking advantage of complementarity in the data. We apply the method on genus profiles derived from three microbiome datasets and we find that CSNF can discover the modular structure of microbial interaction network which cannot be identified by analyzing a single dataset.

species interaction, metagenome, diffusion process, biological network, modularity

Citation: Jiang XP, Hu XH. Inferring microbial interaction networks based on consensus similarity network fusion. Sci China Life Sci, 2014, 57: 1115−1120, doi: 10.1007/s11427-014-4735-x

The interaction among species is important to understand ecological system function. In macroscopic ecology, the structure of food web has been shown to play a pivotal role in the evolution of the environment [1]. However, the structure of inter-species network in microscopic ecology is hard to decipher due to the complexity of microbial community [2]. Until the last decade, the development of high-throughput metagenomic sequencing [3] and 16S sequencing technologies [4] allows the inference of large-scale interactions among microbial species [5].

Computational methods were based on the co-occurrence pattern or correlations of microbial classifications (taxa group in different taxonomic levels, such as species, genus and phylum). Similarity or regression based methods are the most commonly used because of their simplicity and feasibility [5]. Chaffron et al. [6] used Fisher's exact test with false discovery rate (FDR) correction of *P*-value to build a large-scale network in a scale of microbial community based on co-occurrence patterns of microorganism. Many kinds of microorganisms associated in network are also evolutionarily similar. They also found a large number of evolutionarily distant species which are connected in the network. Zupancic et al. [7] used Spearman correlation coefficient to construct a network of relationships among intestinal bacteria from Amish people and used regression analysis to investigate the relationships between three bacterial networks (corresponding to the three enterotypes) and metabolic phenotype, they found 22 bacterial species and four operational taxonomic units (OUTs) related to metabolic syndrome. Faust et al. [8] analyzed the data from the first stage of Human Microbiome Project—16S rRNA se-

*Corresponding author (email: huxiaohua@mail.ccnu.edu.cn)

quence data of 239 individuals with 18 body positions (totally 4302 samples) to build a global interaction network of microorganisms. Friedman et al. [9] systematically investigated the species (or functional) diversity effect on the compositional data and found that the composition effect can be ignored, and vice versa when the data has a very low diversity of species. They proposed a new computational framework (sparse correlation for compositional data, SparCC) to overcome the compositional bias in correlation analysis of microbial co-occurrence data. Tong et al. [10] investigated 179 endoscopic lavage samples from different intestinal regions in 64 subjects analysis of weighted co-occurrence network reveal microbial modules.

These studies indicate that the inference of microbial interaction network is important to understand the structure of a microbial community and hence, its functions and principles adapting its complex inhabit environments. Some other researchers started to infer species pairwise interactions such as competitive and cooperative interactions leveraging to a large-scale microbiome data including high-throughput metagenomes and microbial genomes data [11,12]. These computational efforts facilitated the discovery of previously unknown principles of species interaction network, and verified the consistency and resolved the contradiction of the application of macroscopically ecological theory in microscopical ecology [13].

However, different high-throughput measurements can only provide part of information of a microbial community [14], thus the microbial networks inferred from different microbiome datasets are far from complete [3]. Furthermore, the integration of microbiome datasets is difficult due to the noisy, heterogeneous, distributed and dynamic properties of microbiome data sources [15]. In this study, we focus on the co-occurrence data and we propose to use a similarity network fusion (SNF) [16] approach to integrate the similarities among microbes which are from different datasets. We modify the SNF approach with a consensus *k*-nearest neighborhoods (*k*-NNs) [17] method instead of using *k*-NN [16,18] directly and we find that the modification improves the performance of SNF. We validate the method on simulation data and on the integration of three human gut microbiome datasets.

1 Methods

1.1 Dataset

We use three datasets from the enterotype study by Arumugam et al. [19,20]. The three datasets are from 33 Sanger sequenced gut samples, 16S pyrosequencing data from 154 American individuals [21] and Illumina-based metagenomics data from 85 Danish individuals [22], respectively. The sequences from the three datasets are summarized at the genus level and we use 69 genera which are present across three datasets. Two microbial attributes are downloaded from NCBI database (http://www-ab2.informatik.uni-tuebingen.de/megan/taxonomy/microbialattributes.zip) and used for network analysis: one categorizes whether a bacteria likes oxygen or not (aerobism), and the other categorizes if the bacteria has motility or not (motility). Because a genus contains many species, we use the most representative microbial attributes for each genus. Actually the microbial attributes data contains other microbial features, we tried some of other attributes in our experiments and we found that motility and aerobism are two features which have close connection to network modularity (see Results).

1.2 Similarity network fusion (SNF)

Firstly we construct networks of 69 genera for each of the three available data types and then we efficiently fuse these into one network that represents the full spectrum of underlying data [16,18].

Suppose we have n genera and m measurements (in this paper, they are Sanger sequencing and two high-throughput sequencing techniques—pyrosequencing and Illumina-based sequencing). A genera similarity network is represented by a graph $G=(V, E)$. The vertices V correspond to the microbial genera, and the edges E are weighted by how similar the genera are. A weight matrix W is used to represent all edges, with W_{ij} indicating the similarity between genus i and j. W is often derived by a Gaussian heat kernel function [23]

$$W_{ij} = \exp\left(-\frac{d_{ij}^2}{\mu\varepsilon_{ij}}\right), \tag{1}$$

where μ is a hyperparameter that can be empirically set, d_{ij} is the Euclidean distance between i and j, and ε_{ij} is used to eliminate the scaling problem by the following definition:

$$\varepsilon_{ij} = \frac{mean\left(d\left(i, N_i\right)\right) + mean\left(d\left(j, N_j\right)\right) + d_{ij}}{3}, \tag{2}$$

where $mean(d(i, N_i))$ is the average value of the distances between i and its neighbors.

After defining W, a normalized weight matrix P could be obtained as follows:

$$P_{ij} = \begin{cases} \dfrac{W_{ij}}{2\sum_{k\neq i} W_{ik}}, & j \neq i, \\ \dfrac{1}{2}, & j = i. \end{cases} \tag{3}$$

The normalization is free of the scale of self-similarity in the diagonal entries and avoids numerical instabilities [16].

To define a kernel matrix which could be used to measure local affinity, Wang et al. [16] used k nearest neighbors (k-NN) method:

$$S_{ij}=\begin{cases}\dfrac{W_{ij}}{\sum_{k\in N_i}W_{ik}}, & j\in N_i,\\ 0, & \text{otherwise.}\end{cases} \tag{4}$$

The k-NN method filters out those low-similarity edges and only keeps those k-nearest neighbors for vertices.

Let $P^{(v)}$ and $S^{(v)}$ be the input similarity matrices from the dataset v. The SNF process is to iteratively update similarity matrix corresponding to each data type as follows:

$$P^{(v)}=S^{(v)}\left(\frac{\sum_{k\neq v}P^{(k)}}{m-1}\right)(S^{(v)})^{\mathrm{T}},\ v=1,2,\ldots,m. \tag{5}$$

This procedure updates the status matrices $P^{(v)}$ each time and generates m parallel interchanging diffusion processes on m networks. If two vertices i and j are similar in all data types, their similarity will be augmented through the diffusion process and vice versa.

The final similarity matrix fusing all data types is defined simply as

$$P=\frac{1}{m}\sum_{v=1}^{m}P^{(v)}. \tag{6}$$

Algorithm 1 Algorithms of Ck-NN to compute the consensus information from k-NN

```
SET C=0;
FOR i from 1 to N do
   SET S_i=KNN(i)
   FOR p from 1 to N do
      FOR q from p+1 do
         IF p∈S_i and q∈S_i THEN
          C_pq=C_pq+1
          C_qp=C_qp+1
         ENDIF
         ENDFOR
      ENDFOR
ENDFOR
```

However, k-NN method tends to include noisy edges for every vertex. Thus we propose to use consensus k-NN to solve the problem. The algorithm for finding Consensus k-NN (Ck-NN) is introduced by Premachandran and Kakarala [17]. Firstly, a consensus matrix C among nodes is constructed by Algorithm 1. The matrix C is also normalized to have row sum 1 to satisfy the criterion of kernel matrix alike the definition of eq. (4).

$$C_{ij}=\begin{cases}\dfrac{C_{ij}}{\sum_{k\neq i}C_{ik}}, & i\neq j,\\ 0, & \text{otherwise.}\end{cases} \tag{7}$$

In the modified SNE, $S^{(v)}$ in eq. (5) will be substituted by matrix $C^{(v)}$ which is the consensus matrix from the dataset v.

1.3 Modularity analysis

We cluster the nodes in the microbial network by spectral clustering to see if they tend to organize in modular [24]. Let L be the normalized Laplacian matrix of a weight graph P and $L=I-D^{-1/2}PD^{-1/2}$, the spectral clustering aims to minimize the objective function as follows,

$$\min_{Q\in R^{nk}} Trace\left(Z'LZ\right),\ \text{ s.t. }\ Z'Z=I.$$

We compute the modularity Q which was defined by Newman et al. [25] following the spectral clustering procedure. For a given partition of a network, Q is defined as

$$Q=\frac{1}{2L}\sum_{ij}\left[P_{ij}-\frac{k_ik_j}{2L}\right]\delta_{c_ic_j}, \tag{8}$$

where k_i is the sum of all edge weights of one vertex i and L is the overall sum of weights. $\delta_{c_ic_j}$ is the indicator function and is equal to 0 when i and j belong to the same module. The modularity Q of a partition is high when the number of intra-module edges is much larger than that for a random partition. Q is strictly less than 1, and takes positive values if there are more edges between vertices of the same type than we would expect by chance, and negative if there are less [26].

2 Results and discussion

We firstly compare the consensus SNF (CSNF) with the original SNF on simulation data. The simulation data contains two datasets. Each has 200 data points which could be grouped into two clusters. We find that SNF with consensus k-NN can obtain highest normalized mutual information (NMI) value of 1 when k is from 8 to 20. However, the best NMI score of the original SNF can only reach 0.96 (Figure 1). When k>20, the NMI score of consensus k-NN decrease soon, this is possibly due to the small size of the network. It is interesting to investigate the effect of k for large-scale network in future.

We then apply the modified SNF on three microbiome datasets (Methods). There are 69 genera in all the three datasets and three similarity matrices constructed using heat kernel. Using the SNF diffusion process described in the methods, we get a similarity network among these genera.

Biological networks are modular which is the one of the main contributors to the robustness and evolvability [27]. To see if the genera network has modular structure, we use spectral clustering to cluster the network vertices and compute the modularity measure Q. We use 1000 randomized networks as random controls [28]. Each randomized network is generated by shuffling the edges randomly (1000 times) in the microbial interaction network while keeping both the degree of every node and the degree distribution of

the network unchanged. The Q values of randomized networks are all around 0, but the value of an actual network is 0.49 (Figure 2). We apply spectral clustering on the unfused weight network from three datasets respectively, we find that none network has the modular structure. This suggests that the data integration may provide a comprehensive map which cannot be identified by separate analysis.

We then investigate if the network structure correlates with microbial attributes. We download two microbial attributes from NCBI. The first one indicates if a bacterium exhibits motility. From Figure 3, we can see that there are two big modules. Only three genera (red color) in the left module have motility and most of the genera with motility are distributed in the right modules. We also notice that four of them play as the module-connector, suggesting that bacteria with motility may play a role in a signal communication.

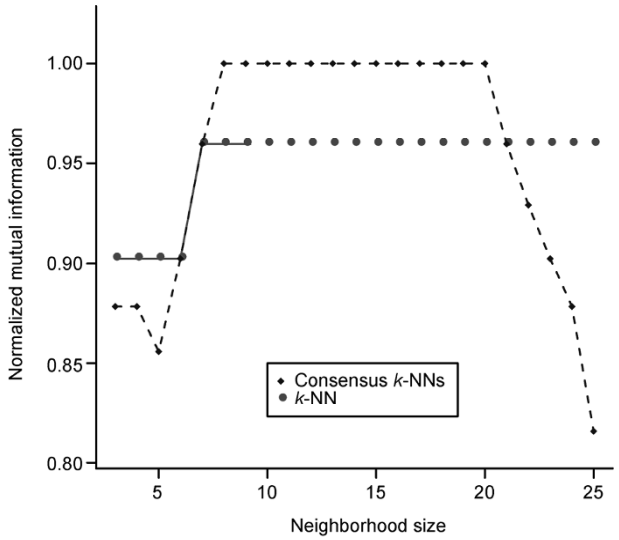

Figure 1 Method comparison on simulation data.

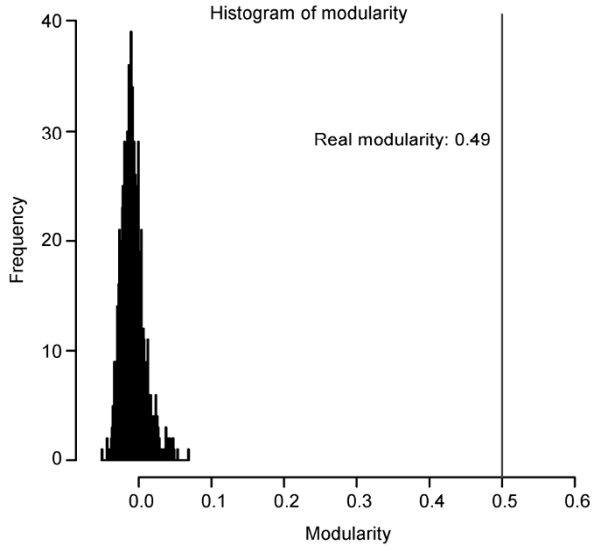

Figure 2 Modularity of the network is significantly higher than that of random network.

The second microbial attribute is to see if the bacteria like oxygen. There are several classifications in this attribute: Aerobic: an aerobic organism is an organism that can survive and grow in an oxygenated environment. Anaerobic: an anaerobic organism does not require oxygen for growth; it may react negatively or even die if oxygen is present. Facultative: an organism can use oxygen but also has anaerobic methods of energy production. Thus it can survive in either environment. Microaerophilic: an organism requires oxygen to survive but requires environments containing lower levels of oxygen than are present in the atmosphere. We label these classifications using different colors in Figure 4 and we find that one module contains mainly anaerobic genera (green). Most of the organisms that like or slightly like oxygen (except two facultative genera) are not in the module. Almost half (6 of 14) of facultative genera tend to be present in another module and half of them tend to play as module-connectors (6 of 14).

These results indicate that the modularity of inferred network correlates with microbial attributes. Thus it seems that microbial organisms with similar microbial attributes tend to have stronger grouping structure than those without. Actually the microbial attributes data contains other microbial features, and we tried some of the other attributes in our experiments and found that motility and aerobism are two features which are closely linked to modularity.

3 Conclusion

A fundamental question in ecology is how different species coexist and interact in natural environments. Microbial species interaction shapes the structure of a microbial community and hence forms its functions and principles adapting to its complex inhabit environments. Deciphering inter-species interaction is hard by wet-lab due to the difficulties of co-culture experiments and the complicate types of species interaction [29]. Although high-throughput experimental approach (e.g., ichip [30]) has developed to detect microbial interactions, it has limited scalability. Development in sequencing technologies and metagenomics now allows researchers to characterize the composition and variation of species across environmental samples, and to accumulate a huge amount of data which provides basis and opportunities to infer the complex principle of species interaction. The construction of microbial network at species level is hard due to the difficulties in obtaining enough data for representing species. This study provides a computational framework for constructing similarity network by integrating similarities among microbial genera from different datasets. In future studies, more efficient similarity fusion methods will be developed for integrating large-scale metagenomic datasets to build a comprehensive microbial interaction network.

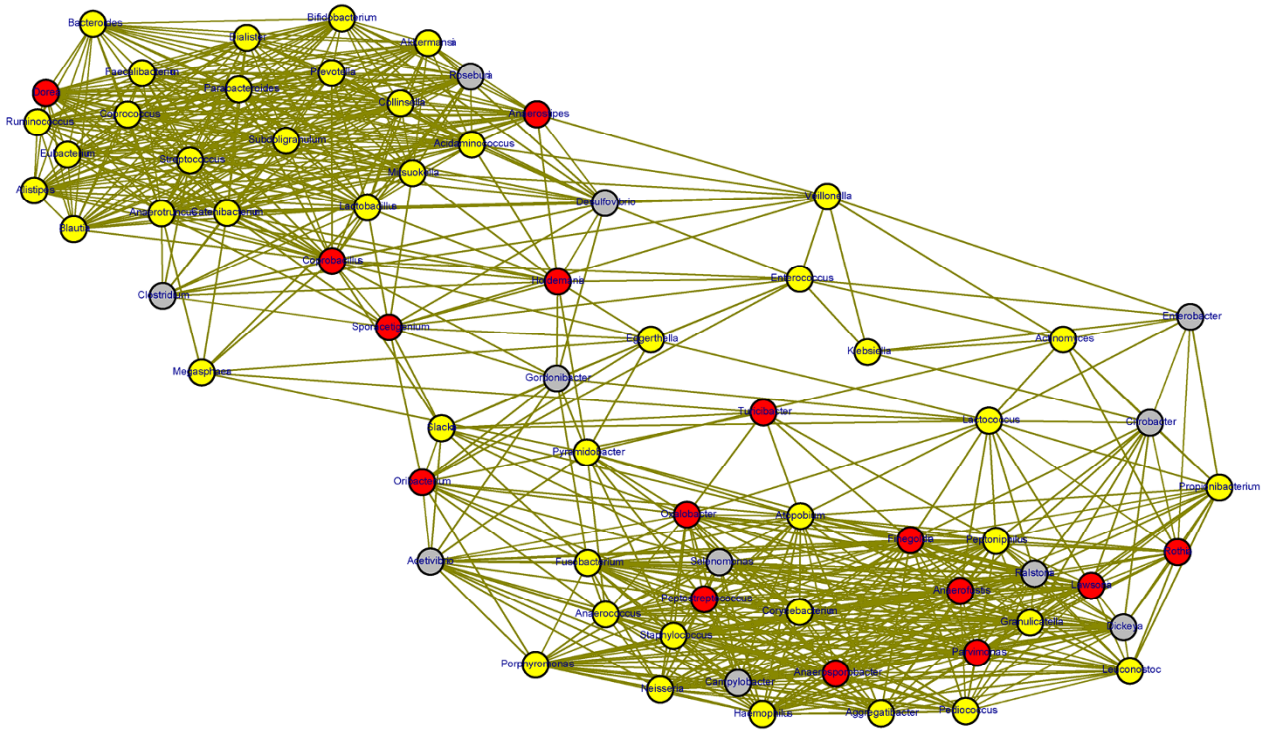

Figure 3 Inferred microbial interaction network: node colors labeled by the motility properties. Red, with motility; yellow, without motility; grey, unknown.

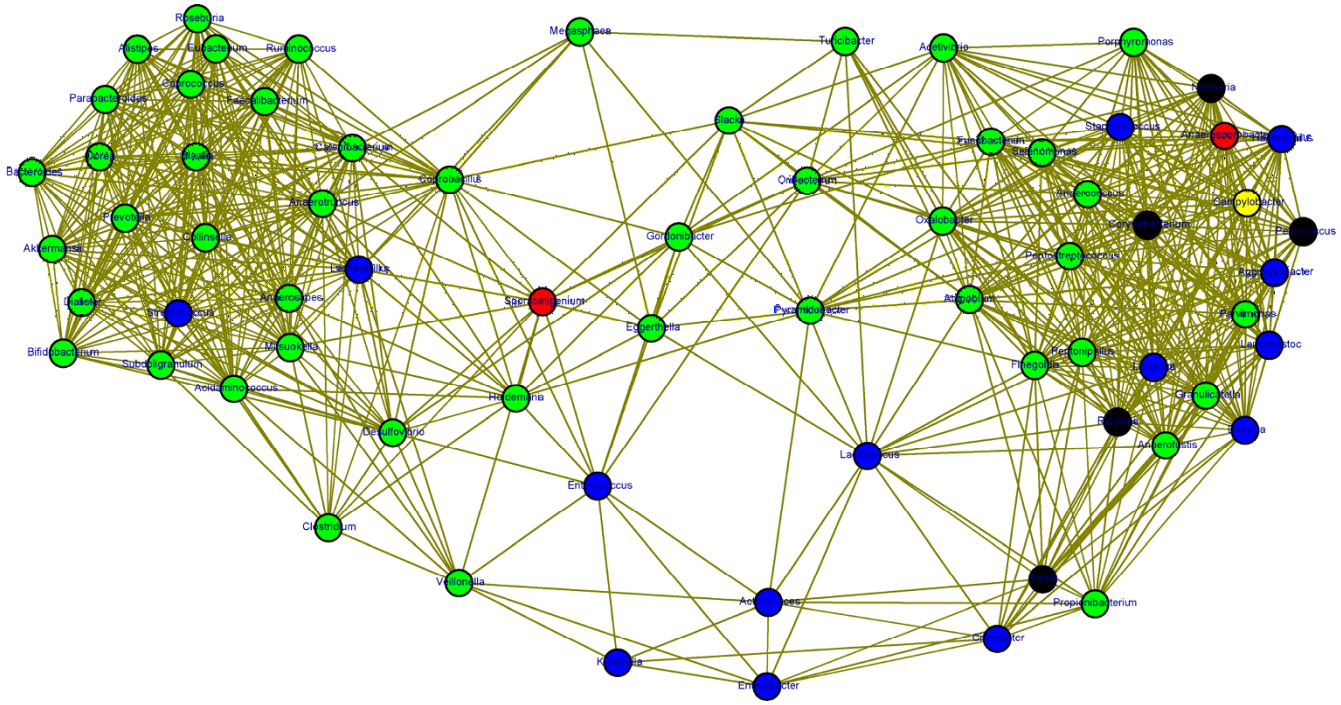

Figure 4 Inferred microbial interaction network: node colors labeled by the oxygen-relate properties. Red, unknown (2); black, aerobic (5); green, anaerobic (47); blue, facultative (14); yellow, microaerophilic (1).

This work was supported in part by US National Science Foundation, Division of Industrial Innovation and Partnerships (1160960 and 1332024), Computing and Communication Foundations (0905291); National Natural Science Foundation of China (90920005, 61170189), the Twelfth Five-year Plan of China (2012BAK24B01), and National Social Science Funds of China (12&2D223, 13&ZD183).

1 Sentis A, Hemptinne JL, Brodeur J. Towards a mechanistic understanding of temperature and enrichment effects on species interaction strength, omnivory and food-web structure. Ecol Lett, 2014, 17: 785–793

2 Chow CE, Kim DY, Sachdeva R, Caron DA, Fuhrman JA. Top-down controls on bacterial community structure: microbial network

analysis of bacteria, T4-like viruses and protists. ISME J, 2014, 8: 816–829
3 Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol, 1998, 5: R245–249
4 Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci USA, 1985, 82: 6955–6959
5 Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol, 2012, 10: 538–550
6 Chaffron S, Rehrauer H, Pernthaler J, von Mering C. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res, 2010, 20: 947–959
7 Zupancic ML, Cantarel BL, Liu Z, Drabek EF, Ryan KA, Cirimotich S, Jones C, Knight R, Walters WA, Knights D, Mongodin EF, Horenstein RB, Mitchell BD, Steinle N, Snitker S, Shuldiner AR, Fraser CM. Analysis of the gut microbiota in the old order Amish and its relation to the metabolic syndrome. PLoS One, 2012, 7: e43052
8 Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C. Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol, 2012, 8: e1002606
9 Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol, 2012, 8: e1002687
10 Tong M, Li X, Wegener Parfrey L, Roth B, Ippoliti A, Wei B, Borneman J, McGovern DP, Frank DN, Li E, Horvath S, Knight R, Braun J. A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease. PLoS One, 2013, 8: e80702
11 Levy R, Borenstein E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proc Natl Acad Sci USA, 2013, 110: 12804–12809
12 Levy R, Borenstein E. Metagenomic systems biology and metabolic modeling of the human microbiome: from species composition to community assembly rules. Gut Microbes, 2014, 5: 265–270
13 Proulx SR, Promislow DE, Phillips PC. Network thinking in ecology and evolution. Trends Ecol Evol, 2005, 20: 345–353
14 Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM. Comparative metagenomics of microbial communities. Science, 2005, 308: 554–557
15 Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol, 2010, 6: e1000667
16 Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods, 2014, 11: 333–337
17 Premachandran V, Kakarala R. Consensus of *k*-NNs for robust neighborhood selection on graph-based manifolds. In: IEEE Conference on Computer Vision and Pattern Recognition, 2013. 1594–1601
18 Wang B, Jiang JY, Wang W, Zhou ZH, Tu ZW. Unsupervised metric fusion by cross diffusion. In: IEEE Conference on Computer Vision and Pattern Recognition, 2012. 2997–3004
19 Siezen RJ, Kleerebezem M. The human gut microbiome: are we our enterotypes? Microb Biotechnol, 2011, 4: 550–553
20 Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Dore J, Antolin M, Artiguenave F, Blottiere HM, Almeida M, Brechot C, Cara C, Chervaux C, Cultrone A, Delorme C, Denariaz G, Dervyn R, Foerstner KU, Friss C, van de Guchte M, Guedon E, Haimet F, Huber W, van Hylckama-Vlieg J, Jamet A, Juste C, Kaci G, Knol J, Lakhdari O, Layec S, Le Roux K, Maguin E, Merieux A, Melo Minardi R, M'Rini C, Muller J, Oozeer R, Parkhill J, Renault P, Rescigno M, Sanchez N, Sunagawa S, Torrejon A, Turner K, Vandemeulebrouck G, Varela E, Winogradsky Y, Zeller G, Weissenbach J, Ehrlich SD, Bork P. Enterotypes of the human gut microbiome. Nature, 2011, 473: 174–180
21 Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. A core gut microbiome in obese and lean twins. Nature, 2009, 457: 480–484
22 Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Jian M, Zhou Y, Li Y, Zhang X, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD. A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 2010, 464: 59–65
23 Lafferty J, Lebanon G. Diffusion kernels on statistical manifolds. J Mach Learn Res, 2005, 6: 129–163
24 von Luxburg U. A tutorial on spectral clustering. Stat Comput, 2007, 17: 395–416
25 Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci USA, 2006, 103: 8577–8582
26 Leicht EA, Newman MEJ. Community structure in directed networks. Phys Rev Lett, 2008, 100: 118703
27 Hintze A, Adami C. Evolution of complex modular biological networks. PLoS Comput Biol, 2008, 4: e23
28 Jiang X, Liu B, Jiang J, Zhao H, Fan M, Zhang J, Fan Z, Jiang T. Modularity in the genetic disease-phenotype network. FEBS Lett, 2008, 582: 2549–2554
29 Lu L, Walker WA. Pathologic and physiologic interactions of bacteria with the gastrointestinal epithelium. Am J Clin Nutr, 2001, 73: 1124s–1130s
30 Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A, Kanigan T, Lewis K, Epstein SS. Use of ichip for high-throughput *in situ* cultivation of "uncultivable" microbial species. Appl Environ Microbiol, 2010, 76: 2445–2450